Information Retrieval – Exercise 4
IST 441

*This exercise is worth **5** points.*
In this exercise, you will create a simplified Lucene index. To get partial credit in case of miscalculations, please give detailed solutions.

Given the following documents:

D1: You say “goodbye,” I say “hello.”
D2: You say “stop”, I say “go.”
D3: “Hello, hello,” you say “goodbye.”
D4: I say “high,” you say “low.”

1. (4 points) Build the inverted index for the documents with tokenization. Do not exclude stop words. Describe what tokenization you use and how you build the dictionary. Build a:

   a. Dictionary file:
      e.g.

Term	**DocFreq**
hello	2
I	3

   b. Posting file (terms are implicit)
      e.g.

Doc #	**Frequency**
1	3
3	3

   c. Position file (terms are implicit from dictionary file, use absolute position of terms in the document)
      e.g.

D1	**D2**	**D3**	**D4**
6,7,8	0	1,2,3	0
4	4	0	1

   d. For a given query

      Q: say goodbye

      Describe and show the process to search the inverted index.

2. (1 points)

    a. Estimate the total size of the inverted index files in bytes. Numbers and characters are counted as 4 bytes. Strings are counted as the number of characters multiplied by 4 bytes. For example, the size of string “hello” is 5*4 = 20 bytes.

    b. Compare the result from 2a. to the total size of the documents in bytes.